Outlier Robust Bayesian Multinomial Choice Modeling

Dries F. Benoit[a,d], Stefan Van Aelst[b,c] and Dirk Van den Poel[a]

[a] Faculty of Economics and Business Administration, Ghent University, Tweekerkenstraat 2, B-9000 Ghent, Belgium

[b] Department of Mathematics, Section of Statistics, KU Leuven, Celestijnenlaan 200B, B-3001 Leuven, Belgium

[c] Department of Applied Mathematics, Computer Science and Statistics, Ghent University, Krijgslaan 281, S9, B-9000 Ghent, Belgium

[d] Corresponding author. Email: Dries.Benoit@UGent.be Phone: +32-9-2643527. Fax: +32-9-2644279

Abstract

A Bayesian method for outlier robust estimation of multinomial choice models is presented. The method can be used for both correlated as well as uncorrelated choice alternatives and guarantees robustness towards outliers in the dependent and independent variables. To account for outliers in the response direction, the fat tailed multivariate Laplace distribution is used. Leverage points are handled via a shrinkage procedure. A simulation study shows that estimation of the model parameters is less influenced by outliers compared to non-robust alternatives. An analysis of margarine scanner data shows how our method can be used for better pricing decisions.

Keywords: choice models, leverage points, outliers, multivariate Laplace distribution, robustness

1 Introduction

Choice models are ubiquitous across a broad range of applications in business and industry. Recent implementations of these models are found in disciplines ranging from revenue management (Farias et al., 2013), over market response estimation (Ebbes et al., 2013), to sales forecasting (Andrews et al., 2013). As one can imagine, important business decisions are based on the estimates obtained from these models. In this paper, a method is proposed

which ensures that these crucial estimates are less biased when small fractions of the data do not adhere to the model assumptions.

Multinomial models relate a qualitative response variable, that consists of $p > 2$ unordered categories to a set of predictor variables. The most well-known models within this class are the multinomial logit model, based on the logistic distribution and the multinomial probit model, based on the multivariate normal distribution (see e.g. Agresti, 2002; Train, 2003).

The multinomial logit and probit models both rely on strong model assumptions and likelihood based estimators are very sensitive to deviations from these model assumptions. In particular, maximum likelihood as well as standard Bayesian estimators for these models can be heavily influenced by a small fraction of contaminated observations in a dataset. This problematic behavior was recognised by Hausman et al. (1998) who proposed a modified likelihood function that deals with misclassification in the dependent variable.

In this paper we consider multinomial models in the context of discrete choice modeling. Multinomial models form a flexible approach to model discrete choice data. In these models the relation between the qualitative choice outcome and the predictor variables is described via an underlying latent continuous variable that is linearly related to the predictors. We focus on Bayesian estimators for multinomial choice models. In this setting we introduce an outlier robust multinomial model that guarantees robustness in both the dependent (i.e. misclassification) as well as in the independent variables (i.e. leverage points).

Misclassification can easily happen when the data is collected in an interview setting where the interviewer makes a mistake in recording the answers. Also in the big-data era, many datasets are messy and noisy, often containing abnormal observations where it is not always clear how they occurred. Other data sources, such as historical datasets can contain measurement errors as well. The method proposed in this study is less vulnerable to these contaminated data points than traditional multinomial models. The model allows for large deviations in the latent responses by using the fat tailed multivariate Laplace distribution (see e.g. Kotz et al., 2001) for the errors in the linear part of the model. As in the univariate case, the center of this distribution can be interpreted as a multivariate median. Hence, by using the multivariate Laplace distribution, we are using a Bayesian method for multivariate median regression.

However, the standard approach of using a heavy tailed error distribution to allow for outlying responses is not sufficient to obtain a fully robust Bayesian estimator that yields reliable results in the presence of outliers and/or leverage points. To account for the possible presence of leverage points, i.e. points that are outlying in the predictor space, following Peña et al. (2009), we allow that the uncertainty on the multinomial model is much larger outside the high density region of the observations and apply a shrinking procedure such that remote observations have much less influence on the posterior and the corresponding Bayesian estimator. Simulations confirm that this shrinkage is of key importance to obtain outlier robust Bayesian inference. To our knowledge this is the first paper that proposes a fully robust Bayesian inference method for flexible multinomial choice modeling.

As is common in Bayesian estimation, exact calculation of the posterior distribution and corresponding Bayesian estimator is not feasible for our model. However, the latent variable formulation that we use for the multinomial model makes this model very attractive for the efficient application of Markov Chain Monte Carlo (MCMC) methods using data augmentation (Tanner and Wong, 1987). Moreover, the representation of the multivariate Laplace distribution as a scale mixture of normal distributions allows to greatly simplify the MCMC procedure for the robust multinomial model.

The rest of this paper is organized as follows. In Section 2 we discuss the general multinomial discrete choice model in detail. The outlier robust Bayesian estimator for this model is introduced in Section 3 while Section 4 explains the computation of the posterior distribution and corresponding Bayesian estimates in the robust multinomial model. The performance of the estimator is investigated through simulations in Section 5. Section 6 contains a real data application and the conclusions and discussion of our approach are summarized in Section 7.

2 The multinomial discrete choice model

We consider multinomial models as methods that relate a set of predictor variables to a qualitative dependent variable consisting of p unordered alternatives where one alternative is chosen by the decision maker. We take the viewpoint that the discrete dependent

variable arises via partial observation of an underlying multivariate continuous variable. This latent variable approach is very flexible. Moreover, in an econometrics context the latent variable can often be given a random utility interpretation which relates latent variable models to the formal econometric specification of demand models based on utility maximization (Rossi et al., 2005).

Consider the general multinomial discrete choice model:

$$\begin{aligned} y_i &= f(u_i); \quad i = 1, \ldots, n \\ \text{with} \quad f(u_i) &= \sum_{j=1}^{p} j \times I\left(\max(u_i) = u_{i,j}\right); \\ \text{where} \quad u_i &= X_i\beta + e_i. \end{aligned} \tag{1}$$

Here, $u_i = (u_{i,1}, \ldots, u_{i,p})'$ is a p-dimensional column vector where p is the number of choice options. Further, $I(\cdot)$ is the indicator function which yields the value 1 if the statement between brackets is true and returns 0 otherwise. In the above notation, $u_{i,j}$ can be interpreted as the *unobserved* utility that is associated with the jth option for unit i. The choice outcome y_i is obtained through the function f which remaps the continuous utility vector u_i onto the discrete space $\{1, \ldots, p\}$. The option that is chosen by unit i, also called the decision maker, is the option with the largest associated unobserved utility $u_{i,j}$. Thus y_i represents the choice among p mutually exclusive alternatives.

The matrices X_i consist of information about the attributes of each of the choice alternatives, as well as covariates which represent the characteristics of the decision making unit. The general structure of the design matrices X_i is $X_i = [b_i' \otimes I_p, A_i]$, where b_i is an r-dimensional vector of characteristics of the decision making unit i and A_i is a $p \times s$ dimensional matrix of choice attributes, and $\otimes$ denotes the Kronecker product. The $k = rp+s$ dimensional vector β contains the regression parameters. Finally, $e_i = (e_{i,1}, \ldots, e_{i,p})'$ is the p-dimensional error vector which follows a distribution F with variance-covariance matrix Ω.

It is well known that for unrestricted Ω model (1) is not identified, i.e. any discrete choice model must be normalized to take account of the fact that the level and scale of utility are irrelevant (see e.g. Train, 1986). A common solution to the former problem is to model the utility differences with respect to a base alternative. For example, if we take

the last alternative as the base choice, then by using the $p-1$ differences $u^\star_{i,j} = u_{i,j} - u_{i,p}$ we can rewrite the model as follows:

$$
\begin{aligned}
y_i &= f(u_i^\star) \qquad (2)\\
\text{with} \quad f(u_i^\star) &= \sum_{j=1}^{p-1} j \times I\left(\max(u_i^\star) = u^\star_{i,j} \;\&\; u^\star_{i,j} \geq 0\right) + p \times I\left(\max(u_i^\star) < 0\right)\\
\text{where} \quad u_i^\star &= D_i\beta + \varepsilon_i\\
\text{and} \quad D_i &= \begin{bmatrix} x'_{i,1} - x'_{i,p} \\ \vdots \\ x'_{i,p-1} - x'_{i,p} \end{bmatrix} \quad \text{where} \quad X_i = \begin{bmatrix} x'_{i,1} \\ \vdots \\ x'_{i,p} \end{bmatrix}\\
\varepsilon_{i,j} &= e_{i,j} - e_{i,p} \quad \text{and} \quad \varepsilon_i \sim G(0, \Sigma).
\end{aligned}
$$

Note that variables that are constant in each choice option (i.e., characteristics of the decision maker) should not be differenced, but can be coded similarly as in Equation 1. That is, these variables are structured in D_i as $[b_i' \otimes I_{p-1}]$.

In Equation 2, the $(p-1)$-dimensional error distribution $G(0, \Sigma)$ has location zero and scatter matrix Σ. The aggregated model for the data can be obtained by stacking the vectors $u_i^\star, \varepsilon_i$ and matrices D_i in a suitable way. Denote $(u^\star)' = ((u_1^\star)', \ldots, (u_n^\star)')$, $\varepsilon' = (\varepsilon_1', \ldots, \varepsilon_n')$ and $D' = (D_1', \ldots, D_n')$, then we obtain:

$$u^\star = D\beta + \varepsilon.$$

To address the arbitrariness of the utility scale, the standard approach to normalize this scale is to normalize the variance of the error terms, i.e. fixing one or more parameters in the scatter matrix Σ. When the within-unit errors can be assumed to be independent and identically distributed, then normalization for scale is straightforward. The variance-covariance matrix of the errors e_i then becomes $\Omega = cI_p$, where I_p is the p-dimensional identity matrix and $c > 0$. In this case normalization is obtained by fixing the constant c. For example, assuming that the independent marginals of F all follow the extreme value distribution and assigning $c = \pi^2/6$, leads to the multinomial logit model (Maddala, 1983).

Assuming that the components of the errors e_i are uncorrelated in model (1) corre-

sponds to the Irrelevance of Independent Alternatives (IIA) assumption (see e.g., Train, 2003) which is often unrealistic and thus undesirable. A model that overcomes this shortcoming and is much more flexible is the multinomial probit model, which emerges when assuming that the errors and consequently the error differences, ε_i, follow a multivariate Gaussian distribution, i.e. $\varepsilon_i \sim N_{p-1}(0, \Sigma)$.

A common approach to solve the scale problem when Σ is not a diagonal matrix is to fix $\Sigma_{11} = 1$. A recent alternative is the trace restriction approach of Burgette and Nordheim (2012). In this paper the latter approach is adopted because it treats all components of Σ equally and can provide stronger identification yielding posterior distributions that are more easily interpreted.

3 Outlier robust Bayesian estimator

In Bayesian statistics, already from the early emergence of the field a lot of research has been conducted to investigate the influence of the prior on the posterior distribution. In contrast to these extensive prior robustness investigations, there is less research about likelihood robustness, i.e. robustness towards outlying data points, in the context of Bayesian estimation. Some exceptions for the univariate linear regression model with a continuous dependent variable are West (1984); Bayarri and Morales (2003); Peña et al. (2009). A Bayesian median regression approach has been proposed by Kottas and Gelfand (2001) in a univariate context and by Dunson et al. (2003) for multivariate regression. In the broader context of quantile regression, Bayesian approaches have been proposed by e.g. Yu and Moyeed (2001); Lancaster and Jun (2010); Taddy and Kottas (2010); Yang and He (2012). For binary outcome variables, a Bayesian quantile regression approach has been introduced by Benoit and Van den Poel (2012). To the best of our knowledge, no outlier robust Bayesian approaches have been proposed for multinomial regression.

Bayarri and Morales (2003) distinguish Bayesian robust methods based on whether or not outlying cases are explicitly identified in the sample. In the former case, different models are fitted to the normal versus outlying cases or, alternatively, the influence of the outlying cases on the likelihood is minimized. In the latter case, a general model is developed, usually with heavy tails, that automatically handles outliers in a suitable way

without requiring their identification. The method proposed in this paper is a combination of both approaches. The latent response is modeled by a moderately heavy tailed distribution, i.e. the multivariate Laplace distribution, that has nice properties in a robustness context. Moreover, extreme outlying observations such as leverage points are identified and their influence on the likelihood function is reduced.

The multivariate Laplace distribution, denoted by MVL(μ, Σ), can be defined through its density function which is given by:

$$f(x;\mu,\Sigma) = \frac{|\Sigma|^{-1/2}}{2^p \pi^{(p-1)/2} \Gamma(\frac{p+1}{2})} e^{-\sqrt{(x-\mu)'\Sigma^{-1}(x-\mu)}} \tag{3}$$

for $x \in \mathbb{R}^p$ and $p \geq 1$. The parameter $\mu \in \mathbb{R}^p$ is the center of the distribution and Σ is the positive definite scatter matrix of size p. The multivariate Laplace distribution is an elliptically symmetric, heavy tailed distribution as illustrated in Figure 1 which shows the density surfaces and contours of two bivariate Laplace distributions. In both cases the center of the distribution equals zero. In the top panel the scatter matrix is I_2, the bivariate identity matrix. In this case the distribution is sperically symmetric. In the bottom panel, the scatter matrix is $\Sigma = \begin{bmatrix} 1 & 3/4 \\ 3/4 & 1 \end{bmatrix}$. We can see the narrow peak around the center in the surface plots on the left of Figure 1. When the dimension of the multivariate Laplace equals one, then the distribution reduces to the well-known univariate Laplace distribution that has been used for Bayesian median regression (see e.g. Yu and Moyeed, 2001; Benoit and Van den Poel, 2012).

If we assume that a sample $x_1, \ldots, x_n$ of independent and identically distributed observations follows the multivariate Laplace distribution in (3), then it has been shown in Arslan (2010) that the corresponding maximum likelihood estimates (MLE) for μ and Σ are the solution $(\hat{\mu}, \hat{\Sigma})$ that minimizes the sum of the distances $d(x_i; m, C) = \sqrt{(x_i - m)'C^{-1}(x_i - m)}$ among all $m \in \mathbb{R}^p$ and positive definite symmetric matrices C of size p with $|C| = 1$. That is,

$$(\hat{\mu}, \hat{\Sigma}) = \underset{m, C, |C|=1}{\operatorname{argmin}} \sum_{i=1}^{n} d(x_i; m, C). \tag{4}$$

The properties of this estimator have been investigated by Roelant and Van Aelst (2007).

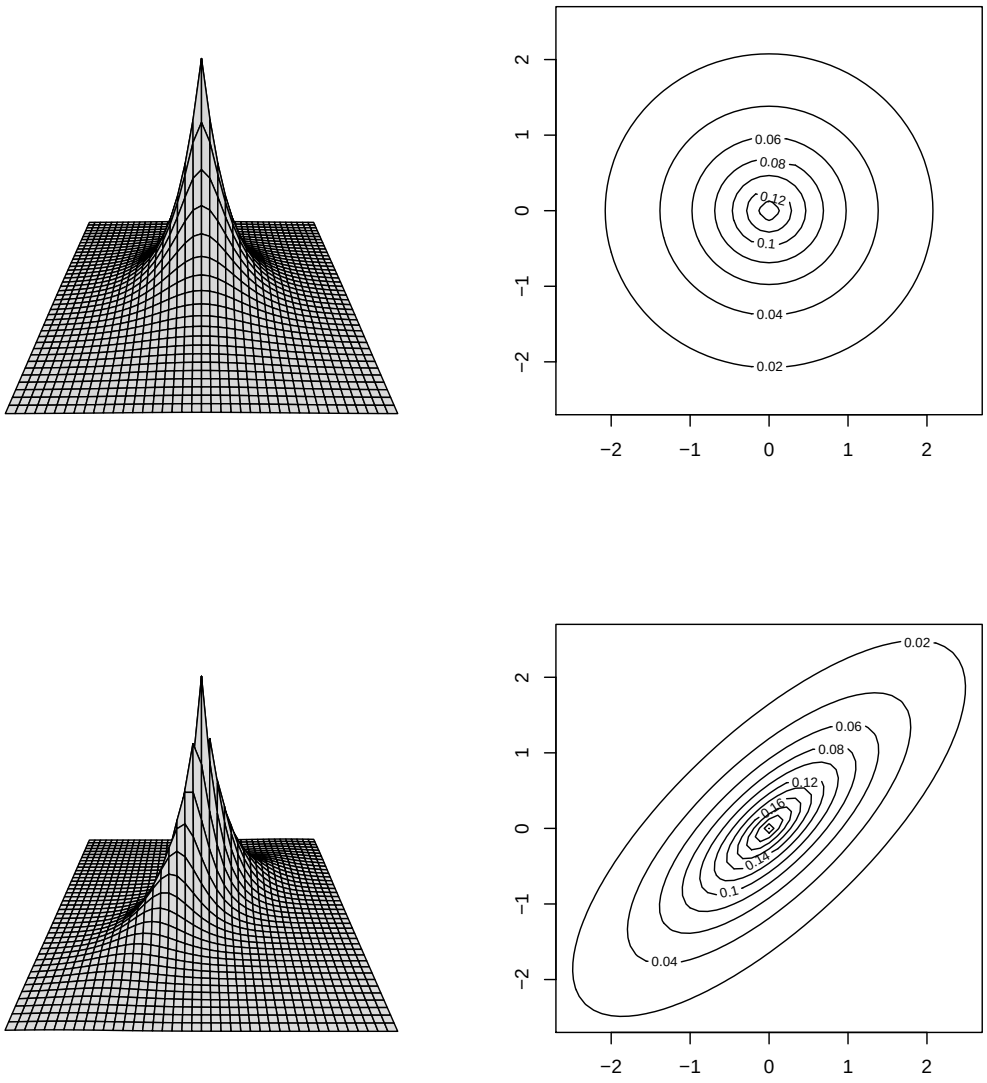

Figure 1: Density surfaces and density contours of the bivariate Laplace distribution $\text{MVL}(\mu, \Sigma)$. The center $\mu = (0, 0)'$ and the diagonal of the scatter matrix Σ equals $(1, 1)'$. Top panel: $\Sigma_{12} = 0$. Bottom panel: $\Sigma_{12} = 3/4$.

Note that for $p = 1$ the estimator $\hat{\mu}$ becomes the univariate L^1 estimator, i.e. the median. Thus, in the multivariate setting $\hat{\mu}$ can be seen as a multivariate extension of the median, obtained by considering the median as the L^1 location estimator. Contrary to other popular affine equivariant multivariate medians, such as the spatial median (see e.g. Brown, 1983), Tukey's halfspace median (Tukey, 1975), or the Oja median (Oja, 1983), the multivariate Laplace MLE comes naturally with an accompanying scatter (shape) estimator. Moreover, since it relies on the MVL, the estimator can be interpreted as a multinomial

median regression technique that has some desirable properties, as we will see later.

Note that modeling the utility differences with a fat tailed distribution is very similar in nature to some of the approaches in the misclassification literature. For example, Hausman et al. (1998) proposed to modify the likelihood function of binary classification models by incorporating two parameters, related to the degree of misclassification of both zeros and ones. The resulting model then becomes:

$$Pr(y_i = 1|x_i) = \alpha_0 + (1 - \alpha_0 - \alpha_1)F(x_i'\beta) \tag{5}$$

Without going into detail, it is immediately clear that the misclassification parameters α_0 and α_1 influence the scale of the link function and consequently the tails of the distribution. Other authors have extended and applied this approach to different situations (see e.g. Ramalho, 2002; Molinari, 2008; Čížek, 2012).

Modeling the conditional response by using a fat tailed distribution such as the multivariate Laplace distribution can only diminish the influence of observations that are outlying in the dependent variable. However, this is not sufficient to obtain a fully outlier robust Bayesian method because the effect of leverage points, is not yet controlled by this approach. To evaluate the robustness of Bayesian inference procedures, Peña et al. (2009) introduced the concept of α-robustness as follows. Consider $\mathcal{Z}$, a sample generated by some parametric statistical model $f(\mathcal{Z}|\theta)$ and let $\mathcal{Z}_\alpha$ denote a contaminated sample that is obtained by replacing a fraction α of points in $\mathcal{Z}$ by arbitrary outliers. Then, Bayesian inference for the parameters θ is called α-robust with respect to the Kullback-Leibler divergence if $\sup \mathrm{KL}(p(\theta|\mathcal{Z}_\alpha), p(\theta|\mathcal{Z})) < \infty$ where the supremum is over all possible contaminated samples $\mathcal{Z}_\alpha$. Here, $p(\theta|\mathcal{Z})$ and $p(\theta|\mathcal{Z}_\alpha)$ are the joint posterior distributions for the parameters θ given the samples $\mathcal{Z}$ and $\mathcal{Z}_\alpha$ respectively.

In Peña et al. (2009) it is shown that Bayesian inference in linear models using a likelihood obtained by assuming i.i.d. errors cannot be α-robust, even when heavy tailed distributions are used. Therefore, additional measures need to be taken to minimize the influence of all types of outliers. To accomplish this, we adopt the procedure of Peña et al. (2009) and extend it to multivariate linear models. Hence, we assume that model (2) holds within the high density region of the observations. Outside this region the uncertainty

on model (2) may be much larger, which is reflected by an increasing scale in the error distribution. Consider the variables $z_i = (b_i', a_{i,1}', \ldots, a_{i,s}')'$ which contain both b_i, the vector of characteristics of the decision making unit i, and $a_{i,1}, \ldots, a_{i,s}$ the vectors of choice specific characteristics (differenced w.r.t. the base alternative) for decision making unit i. Then, for points z outside the high density region the error distribution $G(0, \Sigma)$ is replaced by $G(0, \Sigma/\omega^2(z))$ with $\omega(z)$ approaching 0 for $\|z\| \to \infty$.

A suitable way to determine the high density region is by using a Mahalanobis type distance to the center of the data. The idea behind this procedure is that the uncertainty increases with the distance from the center of the data and thus, remote observations should have less influence on the posterior. To reliably characterize the high density region, the Mahalanobis type distance needs to be based on estimates of the center and scatter of the data distribution that are robust in the sense that they remain reliable even when outlying observations are present in the data. As in Peña et al. (2009) we use computationally cheap estimates for this purpose. That is, for the location we use the coordinatewise median m of z. Hence, the elements of m are the univariate medians of the components of z, i.e. $m_i = \text{med}(z_i)$. Consider the diagonal matrix D whose diagonal elements are the median of absolute deviations with respect to the median (MAD), $\text{MAD}(z_i)$ of the components of z. Moreover, let R be the quadrant correlation matrix (see e.g. Huber, 1981) whose elements are the pairwise quadrant correlations $\text{cor}_{\text{quad}}(z_i, z_j)$ between the components of z. Note that the quadrant correlation between two univariate variables z_i and z_j is the standard Pearson correlation between the pair $\text{sign}(z_i - m_i)$ and $\text{sign}(z_j - m_j)$. Then, the robust estimate of scatter is given by the matrix $C = DRD$.

We now use the Mahalanobis distance of the observations z_i w.r.t. the center m and scatter C of the data, i.e.

$$d_i = d(z_i; m, C) = \sqrt{(z_i - m)' C^{-1} (z_i - m)}. \tag{6}$$

The observations that lie in the high density region satisfy $d_i \leq a$, where a is a well-chosen constant. The constant a determines the robustness of the procedure. To guarantee that the Bayesian inference can withstand a fraction α of leverage points, a can be taken equal to the $(1-\alpha)$-quantile of the distribution of the observation distances d_i. As is common in

robust statistics, choosing the fraction α is a trade-off between robustness and efficiency of the procedure. If α is selected too small, then leverage points can heavily influence the Bayesian estimator, leading to a loss of robustness. On the other hand, a large fraction α leads to more robustness (more shrinkage toward the center), but implies a loss of information if regular observations fall outside the narrowly defined high density region. In the simulations and example of this paper, we let a correspond to the 95%-quantile of the observation distances d_i, i.e. $\alpha = 5\%$.

Finally, to reflect the uncertainty about model (2) outside the high density region, the weights $\omega(z_i) \equiv \omega_i$ are assigned to the observations as follows

$$\omega_i = \begin{cases} 1 & \text{if } d_i \leq a \\ (1 + d_i^2 - a^2)^{-1/2} & \text{otherwise,} \end{cases} \tag{7}$$

with d_i given by (6). The weights ω_i can thus be interpreted as shrinkage factors that pull outlying observations z_i towards the center of the data distribution.

By using the multivariate Laplace distribution in (3) for the errors in model (2) and calculating the weights (7) on the differenced observations, the likelihood of the model, conditional on the latent responses $u_i^\star$ becomes

$$\begin{aligned} L(\beta, \Sigma \,|\, u_i^\star, D_i, \omega_i) &\propto \prod_{i=1}^{n} \omega_i^{p-1} |\Sigma|^{-1/2} e^{-\sqrt{(u_i^\star - D_i\beta)' \omega_i^2 \Sigma^{-1} (u_i^\star - D_i\beta)}} \\ &\propto |\Sigma|^{-n/2} \prod_{i=1}^{n} e^{\sqrt{(\tilde{u}_i^\star - \widetilde{D}_i\beta)' \Sigma^{-1} (\tilde{u}_i^\star - \widetilde{D}_i\beta)}} \\ &\propto L(\beta, \Sigma \,|\, \tilde{u}_i^\star, \widetilde{D}_i) \end{aligned} \tag{8}$$

with $\tilde{u}_i^\star = \omega_i \, u_i^\star$ and $\widetilde{D}_i = \omega_i \, D_i$. This equivalence simplifies the method because the robust Bayesian inference procedure reduces to the usual Bayesian inference for the multivariate linear model (using the multivariate Laplace error distribution) based on the transformed observations. The computation of these Bayesian estimates is discussed in detail in the next section.

It can be argued that the proposed robust Bayesian model is not a genuine Bayesian modeling approach because the weights ω_i are not determined within the Bayesian frame-

work, but are derived from the data beforehand. Recall that the weights are introduced to provide robustness against leverage points and therefore need to be based on robust estimates of the center and scatter of the explanatory variables. Contrary to heteroscedasticity weights, it is not straightforward to include the estimation of these weights in the Bayesian framework through the use of a suitable likelihood.

4 Computation

Consider the multinomial model (2) where the error distribution $G(0, \Sigma)$ is the multivariate Laplace distribution in (3). Moreover, the weight function (7) shrinks the most distant observations towards the center of the data cloud. Note that since the weights are a function of the covariates, we do not condition on them explicitly. Conditioning on the data, $y = (y_1, \ldots, y_n)'$ and $D = (D_1', \ldots, D_n')'$, we denote the full posterior by $\zeta(u^\star, \beta, \Sigma \,|\, y, D)$. To estimate this posterior distribution, we propose a Markov Chain Monte Carlo (MCMC) sampler that cycles through the following conditional distributions:

$$\begin{aligned} u^\star &\mid \beta, \Sigma, y, D \\ \beta &\mid \Sigma, u^\star, D \\ \Sigma &\mid \beta, u^\star, D \end{aligned} \tag{9}$$

This sampler produces a chain with $\zeta(u^\star, \beta, \Sigma \,|\, y, D)$ as its stationary distribution (Tierney, 1994). We can simply marginalize out $u^\star$ if the posterior of β and Σ is desired. Note that there is no need to condition on the response y in steps two to four of the sampling scheme because y does not contribute any extra information once $u^\star$ is known from the first step. Indeed, y is completely determined from $u^\star$ as can be seen from (2).

In the first step, the weights ω_i, as defined in (7), have to be calculated to construct the shrunken sample $\widetilde{D}_i = \omega_i D_i$. Also, to initialize the MCMC procedure we set the starting values for the parameters β and Σ equal to their prior modes. All values of the latent variable $u^\star$ are initially set equal to zero.

Secondly, values of the latent variable $u^\star$ need to be drawn from its distribution conditional on the other parameters and the data. However, the conditional distributions

$u_i^\star \,|\, \beta, \Sigma, y_i, \widetilde{D}_i$ are truncated multivariate Laplace distributions where the truncation on $u_i^\star$ follows from the representation in (2) and can be expressed as

$$\begin{cases} \text{If } y_i = j < p \quad \text{then } u_i^\star \text{ must satisfy} \quad \max(u_i^\star) = u_{i,j}^\star > 0 \\ \text{If } y_i = p \qquad \text{then } u_i^\star \text{ must satisfy} \quad \max(u_i^\star) < 0. \end{cases} \tag{10}$$

Hence, depending on the value of y_i the corresponding $u_i^\star$ must fall within an appropriate cone of $\mathbb{R}^{p-1}$. This truncation makes direct draws from the conditional distribution very difficult to accomplish efficiently.

Drawing from these conditional distributions can be heavily simplified by exploiting that the multivariate Laplace distribution can be represented as a scale mixture of normal distributions (see e.g. Kotz et al., 2001, Theorem 6.3.1, p. 246). This property can be formulated as follows. Consider two independent variables Z and V, where Z is a multivariate standard normal variable, i.e. $Z \sim N_p(0, I_p)$ and V follows an inverse gamma distribution $\text{IG}(\alpha_1, \alpha_2)$ with $\alpha_1 = (p+1)/2$ and $\alpha_2 = 1/2$. Then, the p-dimensional random variable

$$Y = \mu + V^{-1/2}\Sigma^{1/2}Z \tag{11}$$

follows a multivariate Laplace distribution $\text{MVL}(\mu, \Sigma)$ with density function given in (2). If we could observe the variable V, then this property implies that the problem of drawing from Y simplifies to drawing from a multivariate normal distribution, i.e. $Y \,|\, V = v \sim N_{p-1}(\mu, \Sigma/v)$. However, since V is unobservable, we have to estimate it from the data. It turns out that the conditional distribution of V given $Y = y$ is an inverse Gaussian distribution $\text{InvGaus}(\gamma, \lambda)$ whose density function is given by:

$$f(v|\,\gamma, \lambda) = \left(\frac{\lambda}{2\pi v^3}\right)^{1/2} \exp\left(\frac{-\lambda(v-\gamma)^2}{2\gamma^2 v}\right), \tag{12}$$

with $\lambda = 1$ and $\gamma = [(y-\mu)'\Sigma^{-1}(y-\mu)]^{-1/2}$ (see e.g. Arslan, 2010).

The representation in (11) implies that, conditional on V drawing samples from a truncated multivariate Laplace distribution reduces to drawing samples from a truncated multivariate normal distribution $\text{truncN}_{p-1}(\mu, \Sigma/v)$. The latter is an often studied problem which can be performed more easily (see e.g. Geweke, 1991; McCulloch and Rossi,

1994).

In detail, samples of the conditional distributions in the first step of the sampler are obtained as follows. Conditional on y_i, $\widetilde{D}_i$, and the last draws of β and Σ, do for every unit i:

(i) Using the current draw $u_i^\star$, draw the mixing parameter v_i from the inverse Gaussian distribution:

$$v_i \sim \text{InvGaus}(\gamma = [(u_i^\star - \widetilde{D}_i\beta)'\Sigma^{-1}(u_i^\star - \widetilde{D}_i\beta)]^{-1/2}, \lambda = 1) \tag{13}$$

(ii) Draw a sample from the truncated multivariate normal distribution given the truncation in (10) by simulating each element of u_i conditional on the other elements of u_i (see e.g. Geweke, 1991; McCulloch and Rossi, 1994):

$$u_i^\star \,|\, V = v_i \sim \text{truncN}_{p-1}(\widetilde{D}_i\beta, \Sigma/v_i)$$

The resulting draw is a realization from a truncated $(p-1)$-dimensional Laplace distribution with center $\mu_i = \widetilde{D}_i\beta$ and scatter Σ, as required.

In the last two steps of the MCMC sampling procedure (9) we require the conditional posterior distributions for the parameters of a multivariate Laplace distributed sample $u_i^\star \sim \text{MVL}(\widetilde{D}_i\beta, \Sigma)$. A standard approach to obtain these distributions would be to apply a Metropolis-Hastings algorithm (Chib and Greenberg, 1995). However, we propose to use a computationally more efficient approach which is again obtained by exploiting the normal scale mixture representation of the multivariate Laplace distribution in (11).

Note that we have already generated values v_i of the inverse Gaussian random variable V in the first step of (9). From (11) it follows that the conditional distribution $u_i^\star \,|\, V = v_i$ is the multivariate normal distribution $N_{p-1}(\widetilde{D}_i\beta, \Sigma/v_i)$. Hence, conditional on v, the problem of determining the conditional posterior distributions for the regression and scatter parameters of the multivariate regression model with multivariate Laplace errors is simplified to the standard problem of obtaining the conditional posterior distributions for the regression and scatter parameters of the SUR model (Zellner, 1962) with multivariate normal errors.

Until now we have defined the sampler in the unidentified parameter space (as proposed by McCulloch and Rossi, 1994). However, as discussed in Section 2, Burgette and Nordheim (2012) showed the advantages of parameter identification by trace-restricting the covariance matrix. Their approach is adopted in the proposed robust multinomial approach. Denote the unconstrained covariance matrix $\widetilde{\Sigma}$. Define $\Sigma = \widetilde{\Sigma}/\tau^2$ where $\tau^2 = tr(\widetilde{\Sigma})/q$ and $q = (p-1)$, taking as prior for $\widetilde{\Sigma}$ an inverse Wishart distribution with parameters $\bar{\nu}$ and $\bar{\Delta}$ and $tr(\bar{\Delta}) = q$. Furthermore, we choose a conjugate prior for β, i.e. we take an independent multivariate normal prior $\beta \sim N_k(\bar{\beta}, \bar{\Lambda}^{-1})$. This procedure leads to an identified variance-covariance matrix Σ which satisfies $tr(\Sigma) = q$.

When useful prior information about the parameters is available, informative priors can be selected. However, in most cases strong prior information is absent. In that case diffuse conjugate prior distributions on β and Σ can be used such that the posterior distributions of the identified parameters are primarily determined by the information in the sample. Such diffuse priors are obtained by taking the scale (determinant) of $\bar{\Lambda}$ to be small and by choosing a small value for $\bar{v}$ in the inverse Wishart prior. With such settings for $|\bar{\Lambda}|$ and $\bar{v}$ the choice for the remaining parameters in the prior distributions is not critical anymore.

Our MCMC sampling procedure can now be summarized as follows.
First calculate the weights (7) and construct the shrunken sample ($\widetilde{D}_i = \omega_i \, D_i$). Then, initialize the parameters:

$$\beta = \bar{\beta} \qquad \widetilde{\Sigma} = (n-p-2)\bar{\Delta} \qquad u_i^\star = 0,$$

and iterate the following steps:

1. For $i = 1, \ldots, n$ generate values from the conditional distributions $u_i^\star \,|\, \cdot$ by

   (i) Drawing values v_i from the inverse Gaussian distributions in (13).

   (ii) Drawing $u_i^\star$ from the truncated multivariate normal distribution $\text{truncN}_{p-1}(\widetilde{D}_i\beta, \Sigma/v_i)$ using the truncation in (10).

2. Draw τ^2 from the scaled inverse chi-square distribution $tr(\bar{\Delta}\Sigma^{-1})/\chi^2_{(\bar{\nu}q)}$.

3. Transform $\tilde{u}_i^\star = \tau u_i^\star$.

4. Draw $\tilde{\beta}$ from its conditional posterior distribution $\tilde{\beta} \,|\, \cdot \sim N(\hat{\beta}, \tau^2 \hat{\Lambda}^{-1})$ with

$$\begin{aligned}
\hat{\Lambda} &= \left(\bar{\Lambda} + \sum_{i=1}^{n} \widetilde{D}_i' \left(\nu_i \Sigma^{-1} \right) \widetilde{D}_i \right), \\
\hat{\beta} &= \hat{\Lambda}^{-1} \left(\bar{\Lambda}\bar{\beta} + \sum_{i=1}^{n} \widetilde{D}_i' \left(\nu_i \Sigma^{-1} \right) \tilde{u}_i^{\star} \right), \\
\tau^2 &= \frac{\sum_{i=1}^{n} (\tilde{u}_i^{\star} - \widetilde{D}_i' \hat{\beta})' \left(\nu_i \Sigma^{-1} \right) (\tilde{u}_i^{\star} - \widetilde{D}_i' \hat{\beta}) + \hat{\beta}' \bar{\Lambda} \hat{\beta} + tr(\bar{\Delta}\Sigma^{-1})}{\chi^2_{(n+\bar{\nu})q}}.
\end{aligned}$$

5. Draw $\widetilde{\Sigma}$ from the inverse Wishart distribution $\widetilde{\Sigma} \,|\, \cdot \sim \text{InvW}(\bar{\nu} + n, \hat{\Delta})$ with

$$\hat{\Delta} = \bar{\Delta} + \sum_{i=1}^{n} \nu_i \, (\tilde{u}_i^{\star} - \widetilde{D}_i' \tilde{\beta})(\tilde{u}_i^{\star} - \widetilde{D}_i' \tilde{\beta})'.$$

6. Set the following variables:

   (i) $\tau^2 = tr(\widetilde{\Sigma})/q$

   (ii) $\Sigma = \widetilde{\Sigma}/\tau^2$

   (iii) $u^{\star} = \tilde{u}_i^{\star}/\tau$

   (iv) $\beta = \tilde{\beta}/\tau$

An advantage of our MCMC procedure is that it avoids the difficult evaluation of choice probabilities when estimating the model parameters. If desired such choice probabilities can be obtained afterwards by integrating the multivariate Laplace distribution over $(p-1)$-dimensional cones. In detail,

$$\Pr(y_i = j \,|\, D_i, \beta, \Sigma) = \int_{R_{i,j}} G(u^{\star} \,|\, D_i, \beta, \Sigma) \, du^{\star},$$

where G is the multivariate Laplace distribution $\text{MVL}(D_i\beta, \Sigma)$. The region $R_{i,j}$ is the set of vectors $u^{\star}$ that correspond to the truncation in (10). That is,

$$\begin{aligned}
&\text{If } j < p, \quad \text{then} \quad R_{i,j} = \{u^{\star} : \max(u^{\star}) = u_j^{\star} \;\&\; u_j^{\star} > 0\} \\
&\text{If } j = p, \quad \text{then} \quad R_{i,j} = \{u^{\star} : \max(u^{\star}) < 0\}
\end{aligned}$$

These integrals can not be evaluated analytically. However, their solution can be approxi-

mated, for example, by means of simulation. The idea is to simulate many random draws from the multivariate Laplace with mean and covariance matrix equal to the estimated mean and covariance matrix and then calculating the proportion of draws that fall in the specific regions we want to integrate over. For more information about approximating integrals with simulation methods see Geweke (2001).

5 Simulations

In this section we discuss the results of simulations that were performed to assess the performance of the proposed outlier robust Bayesian inference for the multinomial model. The robust Bayesian estimates for the identified parameters are obtained as the mean of their posterior distribution. In the first simulation design we examine the performance of the estimator when the data are actually generated from the assumed model. This allows us to evaluate the effectiveness of the inference, i.e. its performance under ideal circumstances. Also, this setting is used to investigate the performance of the proposed MCMC sampler. In the second set of simulations we then generate the data from alternative models to evaluate the robustness of the estimator.

5.1 Effectiveness

We first investigate the performance of the proposed estimator when the data follow the ideal model. Therefore, we generated $n = 500$ observations according to model (2) with $p = 3$ and $k = 4$. For the design matrices we use $D_i = [I_2 \; A_i]$ with A_i a 2×2 matrix whose elements are generated independently from a uniform distribution on the interval $(0, 2)$. The error distribution $G(0, \Sigma)$ is the multivariate Laplace distribution. The parameters are set equal to the following values (note that $tr(\Sigma) = (p - 1)$ which corresponds to the identification restriction in the MCMC sampler):

$$\beta = (-0.5,\, 0.75,\, 0.75,\, 0.25)' \qquad \text{and} \qquad \Sigma = \begin{bmatrix} 1 & 0.3 \\ 0.3 & 1 \end{bmatrix}.$$

To calculate our robust Bayesian estimator, the MCMC algorithm is run with 5000 iterations starting from diffuse priors. That is, for β we used independent normal dis-

tributions centered at zero with a variance of 100. For Σ we used a Wishart prior with $\bar{v} = p + 3$ degrees of freedom and $\bar{\Delta} = I_{p-1}(p - 1 + 3)$. This prior is vague, but still informative enough to be able to set up a stable MCMC sampler. The results for the estimation of the regression coefficients β are presented in Figure 2 and in Figure 3 for the scatter parameter Σ.

The plots on the left side of Figures 2 and 3 show the posterior distributions for the coefficients β and the elements of the scatter matrix Σ respectively. The vertical dashed lines in these plots indicate the true parameter values. The shaded area under the posterior curves denotes the posterior 95% credible interval. For both the regression coefficients and the covariance matrix, our estimation procedure shows a good performance. All credible intervals contain the true parameter value. Moreover, the mass of all posterior distributions for the regression coefficients is located away from zero, which confirms the relevance of all covariates in the multinomial model.

The right hand side of Figures 2 and 3 show the trace plots of these model parameters. Note that the MCMC chains were not thinned nor was any burn-in excluded from the plots. The trace plots show good mixing properties of the MCMC algorithm. The initial values of β and Σ wear off very fast and the chain then navigates nicely through the parameter space.

However, the trace plots also indicate that the sampler might exhibit some autocorrelation. A more formal investigation is provided by the autocorrelation function (ACF) in the left panel of Figure 4. We only show the ACF plot for the first regression coefficient β_1, but this plot is representative for all other model parameters. The plot indicates that the sampler indeed exhibits quite strong autocorrelation. Even after 35 draws the effect of the first draw is still present. This behavior is typical for latent variable models (Rossi et al., 2005). Fortunately, autocorrelation is not a crucial issue since the MCMC chains still converge to the true posterior distribution as long as a sufficient number of posterior draws are sampled. It does imply that the computation time increases because a large enough number of iterations should be run in the chains in order to fully travel the space of the posterior distribution. The ACF plot for such a long chain is given in the right panel of Figure 4. This plot shows that the problem of autocorrelation disappears for longer thinned MCMC chains.

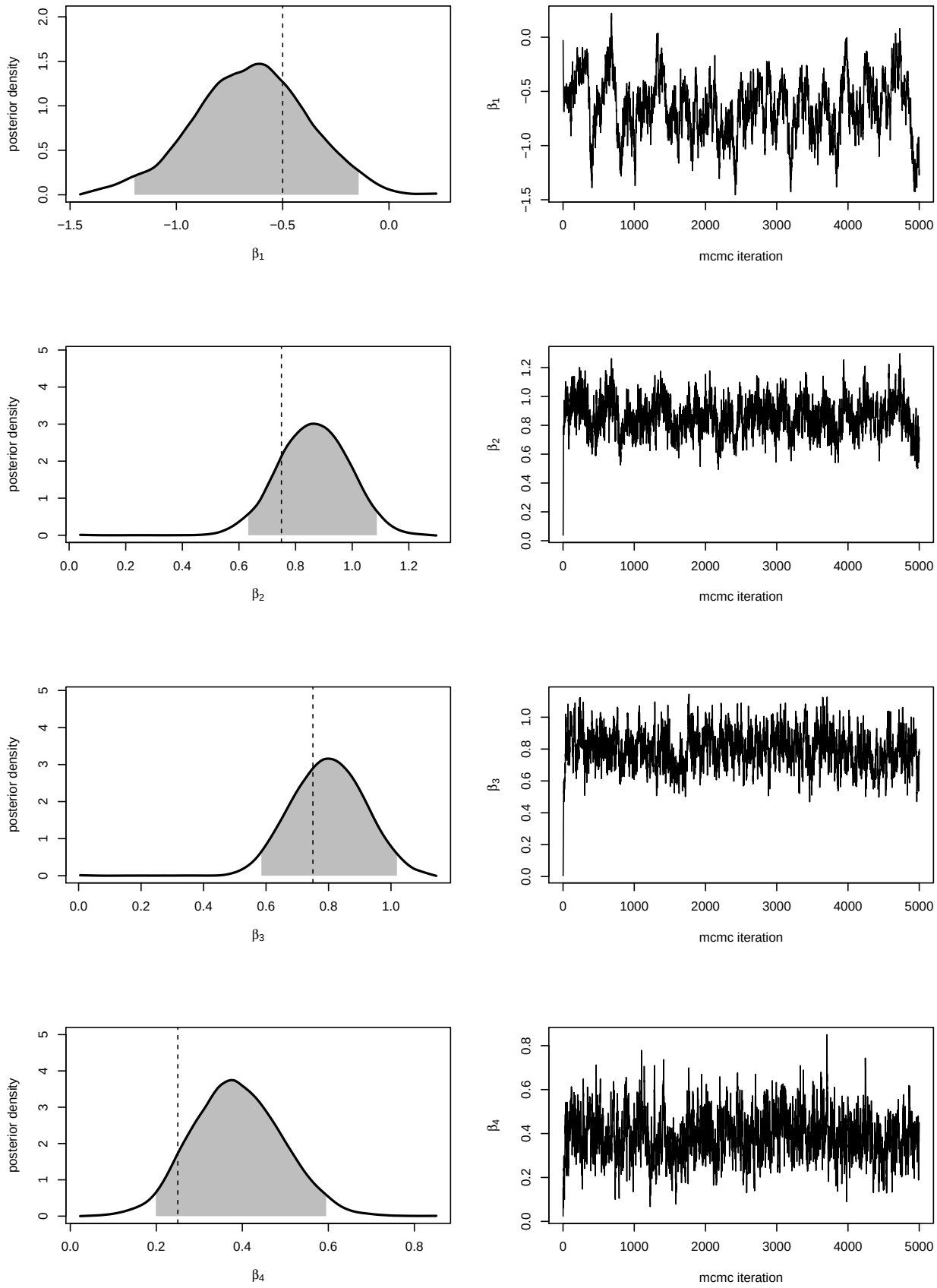

Figure 2: Posterior distributions and trace plots of β.

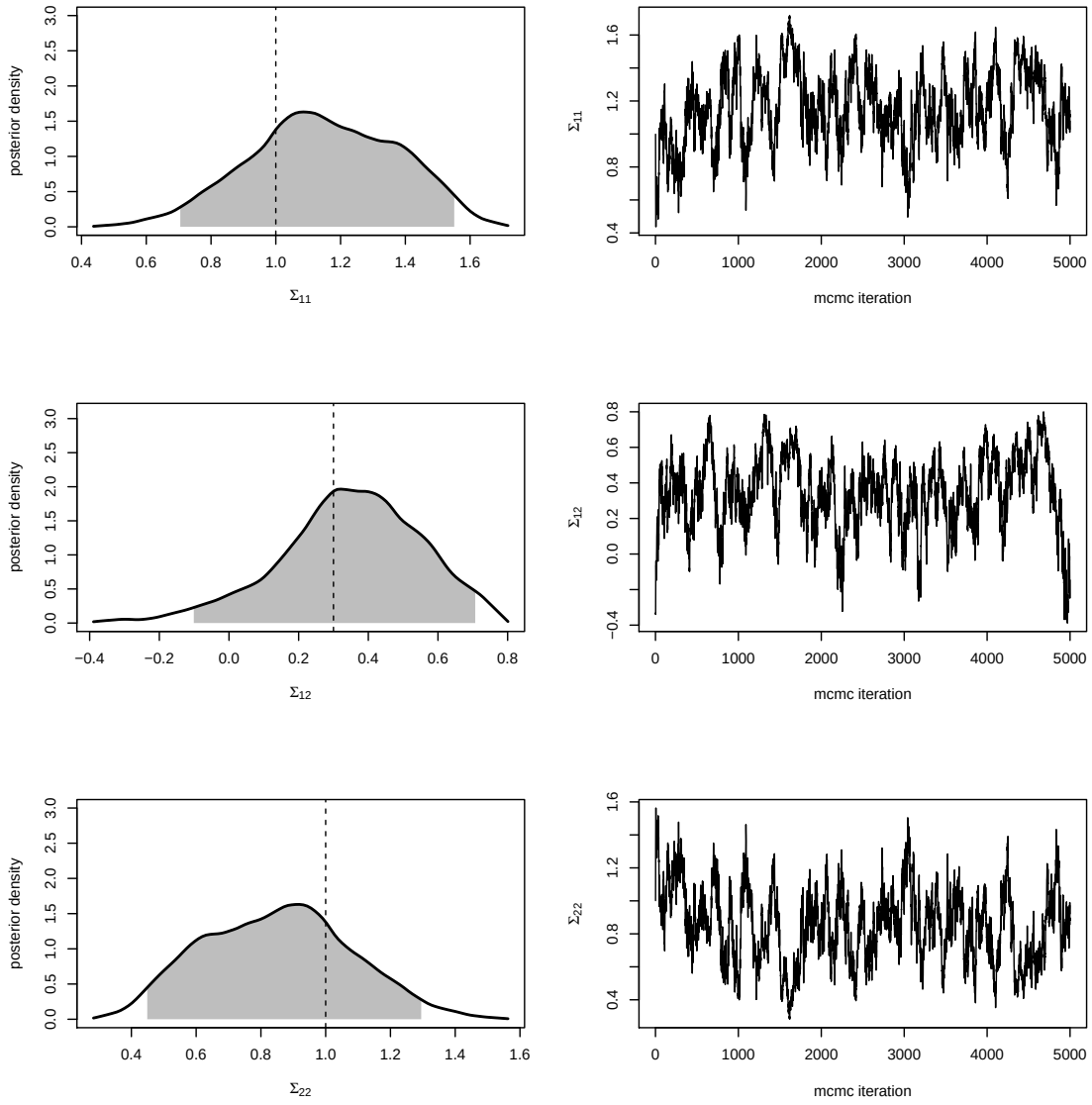

Figure 3: Posterior distributions and trace plots of the three unique elements of Σ.

From this simulated example we can conclude that the proposed robust Bayesian procedure is able to retrieve the parameters that governed the data generating process as desired.

5.2 Robustness

In these experiments, an extensive comparison of the proposed outlier robust Bayesian multinomial method (RMN) with the standard multinomial probit method (MNP) is conducted. For this comparison, we simulated data from 7 different data generating pro-

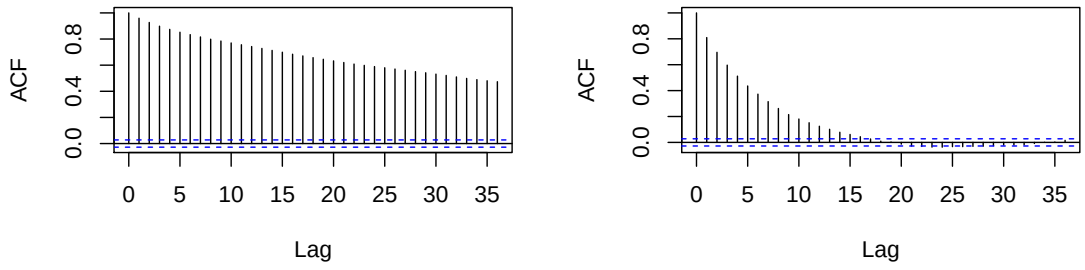

Figure 4: Autocorrelation for β_1 in MCMC sampler. Left panel: 5000 draws without thinning. Right panel: 50,000 draws keeping every 10^{th} draw.

cesses. For each data generating mechanism, datasets with three different sample sizes ($n = \{100, 200, 400\}$) are simulated and both the RMN and MNP method are applied to the datasets. This procedure is then replicated 1000 times. Finally, for every sample size the bias and root mean squared error are calculated for both methods.

The first data generating process follows model (2) with $p = 3$ and $k = 3$. The design matrices D_i are vectors of length 3 whose elements are generated independently from a uniform distribution on the interval $(0, 2)$. The error distribution $G(0, \Sigma)$ is now taken to be the multivariate normal distribution. Moreover, the parameters are set equal to the following values

$$\beta = (0.5, 1, -1)' \qquad \text{and} \qquad \Sigma = \begin{bmatrix} 1 & 0.3 \\ 0.3 & 1 \end{bmatrix}.$$

These parameter settings result in a balanced choice vector, i.e. the unconditional response probabilities for every option are close to 1/3. Also, $tr(\Sigma) = (p - 1)$ which corresponds to the identification restriction in the MCMC sampler. The resulting datasets thus meet all assumptions of the multinomial probit model. We refer to datasets generated according to this model as CLEAN data because there are no deviations from the multinomial probit model that might upset the MNP estimator. Based on this clean data generating model we then consider 6 modifications that contaminate the datasets with different amounts and types of outliers. This allows us to investigate the robustness properties of the Bayesian estimators.

The first two data generating processes, called RESP1 and RESP2, contaminate 5%

and 10% of the clean data, respectively, by misclassifying the responses. That is, instead of choosing the category with the highest latent utility, the option with the lowest latent utility is chosen:

$$f(u_i^\star) = \sum_{j=1}^{p-1} j \times I(\min(u_i^\star) = u_{i,j}^\star \;\&\; u_{i,j}^\star < 0) + p \times I(\min(u_i^\star) > 0)$$

This approach induces contamination only in the response direction and has also been used in the context of binary choice models (see e.g. Čížek, 2012).

The next two data generating processes, called LEV1 and LEV2, contaminate 5% and 10% of the data respectively by simulating the regressor D_i from a different distribution, i.e. $D_{ij} \sim N(-0.5, 16)$ for $j = 1, 2, 3$. The corresponding responses are still generated according to the multinomial probit model. Hence, these observations are outliers in the covariate space, i.e. leverage points, but they still completely adhere to the assumptions of the multinomial probit model. Such observations are considered to be *good leverage* points because they do not bias the results but instead facilitate estimation of the model parameters.

Finally, the last two data generating processes, called OUTL1 and OUTL2, contaminate 5% and 10% of the data by generating leverage points as in the previous case. Moreover, for these leverage points the response is contaminated as well by choosing the option with the lowest latent utility as before. Hence, observations that are outlying in both the response and predictor space are obtained, i.e. outlying observations in the predictor space combined with misclassification. Due to the high variance of the contaminated predictors, these outliers, also called *bad leverage* points, can have a high leverage effect on non-robust model fits. This means that non-robust estimators of the linear model are attracted by the outliers and thus such estimates are heavily biased by these outliers (see e.g. Mebhane and Sekhon, 2004).

To calculate the Bayes estimators (i.e. the expected value of the posterior distribution of the model parameters) for both the MNP and RMN models, the MCMC algorithm is run with 10,000 iterations starting from the modes of the same proper diffuse priors as in the previous simulation. For both methods, the first 2,000 iterations were discarded as burn-in draws and a thinning factor of 10 was used to decrease autocorrelation in the

MCMC chains.

The results of these Monte Carlo experiments are summarized in Figures 5 and 6. These figures contain dot charts that represent both the bias and root mean squared error (RMSE) for each of the simulation settings, averaged over all regression parameters.

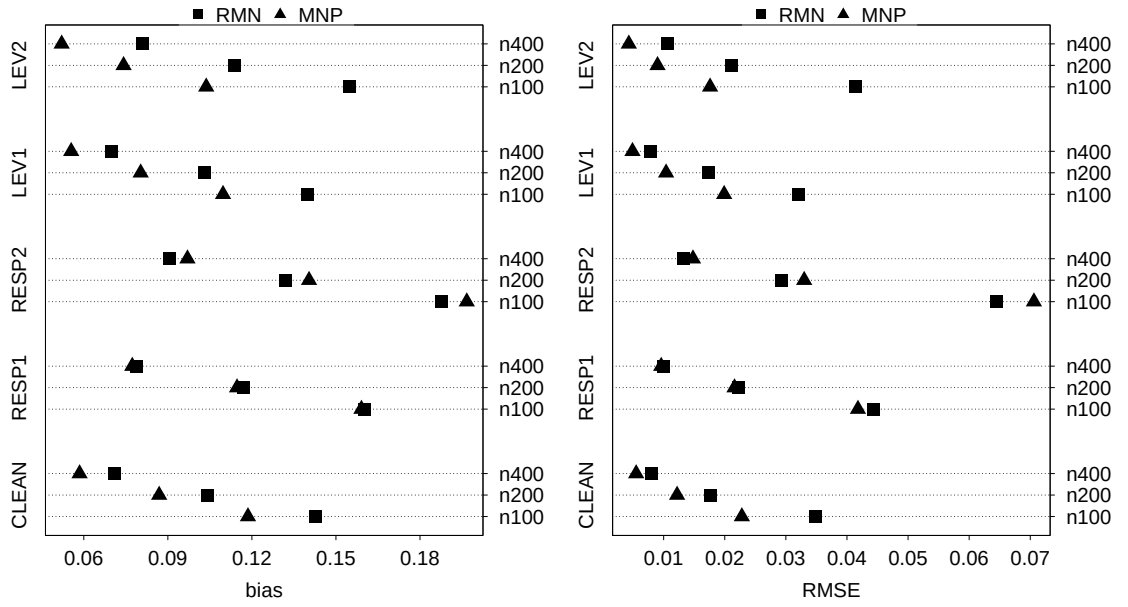

Figure 5: Bias (left panel) and root mean squared error (right panel) of both the proposed Robust Multinomial model (RMN) and standard Multinomial probit model (MNP) for the different sample sizes (n={100,200,400}) and data generating processes: no outliers (CLEAN), 5% misclassification (RESP1), 10% misclassification (RESP2), 5% good leverage points (LEV1) and 10% good leverage points (LEV2).

Figure 5 shows that both RMN and MNP perform best on the CLEAN dataset and the datasets with *good leverage* points. On these datasets that adhere all MNP assumptions, the MNP model has lower bias and RMSE than the proposed robust approach, as expected. The difference is larger in the presence of good leverage points due to the shrinkage of these points in the RMN model which slightly reduces its performance.

Outliers in the response direction (RESP1 and RESP2) affect the performance of both models. The differences between MNP and RMN are rather small for modest amounts of outliers of this type. For higher levels of outliers in the response direction, the advantages of the robust approach become more pronounced. Also note that both models perform better when more data become available, as expected.

Figure 6 shows the bias and RMSE for both methods on datasets with *bad leverage* points in comparison to their performance on CLEAN data. Note that the scale of the

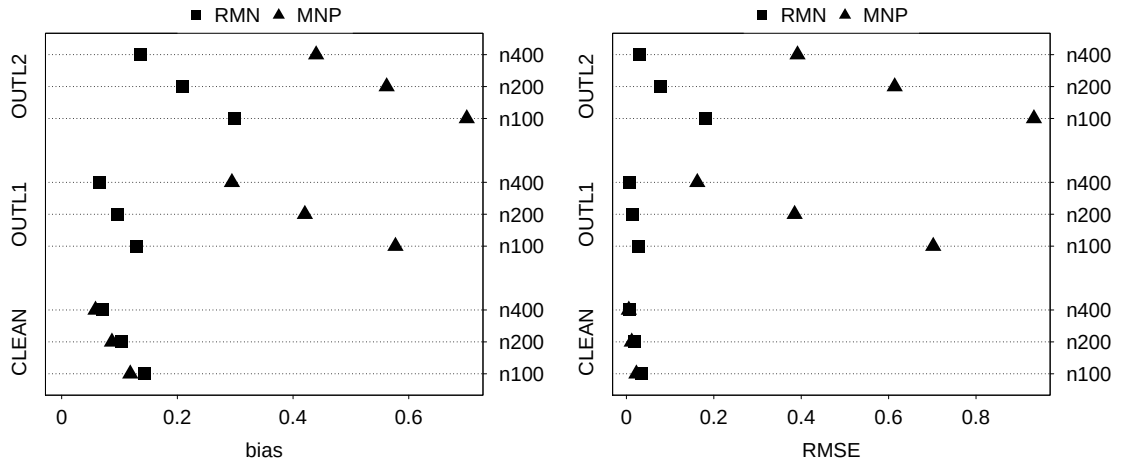

Figure 6: Bias (left panel) and root mean squared error (right panel) of both the proposed Robust Multinomial model (RMN) and standard Multinomial probit model (MNP) for the different sample sizes (n={100,200,400}) and data generating processes: no outliers (CLEAN), 5% bad leverage points (OUTL1) and 10% bad leverage points (OUTL2).

plots in Figure 6 is much larger than the scale in Figure 5. The results clearly show that our RMN approach outperforms the standard MNP method for both levels of outlier contamination. The differences between MNP and RMN in Figure 5 become futile when compared with the differences shown in Figure 6. Remember that in RMN we fixed the constant a in (7) such that the 5% most outlying cases are shrunk toward the center of the data cloud. This explains the good behavior of this approach up to 5% of contamination.

For the designs with 10% of contaminated observations, our choice of α only shrinks the largest 5% of outliers toward the center of the data. The remaining outliers can not be downweighted by the shrinkage procedure. Hence, with 10% of bad leverage points the effect on the RMN procedure is much larger (but still not as large as the effect on the MNP model) because the leverage effect of all outliers can not be downweighted anymore. This illustrates the need to make a good, i.e. large enough, choice of α, such that all leverage points can be shrunk toward the center of the data. We suggest to perform a sensitivity analysis with regard to the fraction of extreme observations that should be shrunk toward the center of the data cloud. That is, re-estimate the model by varying the fraction α over a sensible range (e.g. 5%-20%).

To illustrate the effect of shrinking a fraction α of most extreme observations, we repeated the simulations with bad leverage points (OUTL1 and OUTL2), where we now also included the common approach of merely using a heavy tailed error distribution to

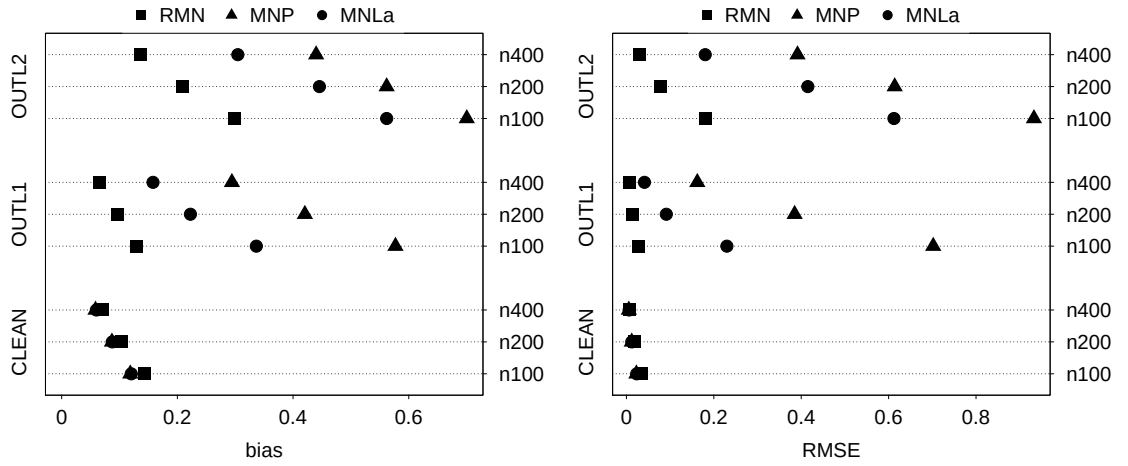

Figure 7: Bias (left panel) and root mean squared error (right panel) of both the proposed Robust Multinomial model (RMN), the standard Multinomial probit model (MNP) and the multinomial Laplace model (MLa) for the different sample sizes (n={100,200,400}) and data generating processes: no outliers (CLEAN), 5% bad leverage points (OUTL1) and 10% bad leverage points (OUTL2).

handle outliers. That is, we fix the weights ω_i equal to one in the MCMC procedure so that no shrinkage is applied and we denote this multinomial Laplace model by MNLa. We know that this approach is not fully robust and now investigate the practical consequences for inference using datasets with bad leverage points. Figure 7 shows the bias and RMSE of the three models.

We can see that the performance of the MNLa model always falls in between the performance of the RMN and MNP model. Clearly, modeling the data with a fat tailed distribution improves the standard MNP model in the presence of bad leverage points, but not to the same extent as in combination with the proposed shrinkage procedure. As expected, this effect is again largest when the chosen fraction of α is large enough to shrink all leverage points. This illustrates the importance of the combination of both the Laplace with the shrinkage procedure to obtain truly outlier robust Bayesian inference.

6 Example

A scanner panel data set of purchases of margarine is used to illustrate the proposed robust multinomial choice model. The complete dataset is comprised of 9196 purchases of ten brands of margarine by 517 households in Springfield, MO. The dataset was extracted from a larger database of household data gathered by A.C. Nielsen company. This dataset

was first analyzed in Allenby and Rossi (1991) and later in Burgette and Nordheim (2012). The dataset was made available through the R-package *bayesm* (Rossi, 2011).

At first it might seem odd to use a robust approach on scanner data as the probability of erroneous data in scanner data is relatively small (but not impossible). However, note that typically scanner data is augmented with loyalty card information such as socio-demographics and these data are much more likely to contain errors. In addition, the proposed model is not only relevant when erroneous data is present, but also when the outlying observations are correct. Typically, researchers do not want their parameter estimates to be influenced by a small number of observations that behave completely different from the bulk of the data. In this dataset, this could be an extremely loyal customer whose price sensitivity is virtually zero. The proposed method will minimize the influence of these outlying (but possibly correct) observations.

To illustrate our approach using this real data example, the panel structure of the data is transformed in a cross-sectional structure by retaining only the first purchase of the household. In this way, we avoid the additional complexity of estimating panel data, but note that our robust RMN model could be extended to analyze this type of data. Moreover, we focus on five brands of margarine: Blue Bonnet, Fleischmann's, House Brand, Generic and Shed Spread.

This results in a dataset with 242 households and an equal number of observed choices. Shed Spread is chosen as the base alternative in our analysis. Two covariates are included in the model. The first variable is Price, measured in US dollar. This is a variable that varies over the choice options. The second variable, log of Household Income also measured in US dollar, is case specific but constant over the choice options. Finally, brand specific intercepts for the four alternative brands are included in the analysis as well, so that we have $k = 9$ regression parameters in the model.

We use the same diffuse prior settings as in the simulation designs in the previous section. The MCMC sampler was ran for 10,000 iterations with a thinning factor of 2. The first 1,000 draws were discarted as burn-in draws. The results of this analysis are presented in Table 1 which shows the robust Bayes estimates (i.e. the expected value of the posterior distributions) and the corresponding 95% credible intervals for the regression parameters in the model.

As in all multinomial choice models, only the coefficients of the variables that vary across the choice options are directly interpretable. In this example, this is only the case for the Price variable. As expected, this variable has a negative sign, meaning that an increase in the price of one alternative decreases the probability of choosing that alternative. The interpretation of the coefficients of the variables, such as Household Income, that do not vary across choice options is less straightforward as the interpretation is relative to the base alternative. A positive sign of a regression coefficient (e.g. Income Fleischmann's= 0.039) does not imply that an increase in the regression variable leads to an increase in the probability of choosing the corresponding alternative. Interpretation is, for example, that compared to Shed Spread (the reference category) a higher household income leads to an increased probability of buying Fleischmann's (since $0.039 > 0$). Similarly, the marginal effects of the Household Income variable could be investigated, but because most of its parameter estimates are close to zero this is not very relevant in this example. Note that the finding that household income does not have an big impact on the choice probabilities confirms the findings of Allenby and Rossi (1991). They argue that this results from the fact that the income variable was measured very imprecise.

Variable	Posterior Mean	95% Credible Interval	
Intercept Blue Bonnet	-2.114	-3.145	-1.122
Intercept Fleischmann's	-1.044	-2.611	0.275
Intercept House Brand	-3.510	-4.865	-2.223
Intercept Generic	-4.619	-6.133	-3.211
Price	-7.218	-9.658	-4.965
Income Blue Bonnet	0.007	-0.024	0.038
Income Fleischmann's	0.040	-0.003	0.085
Income House Brand	0.016	-0.023	0.053
Income Generic	0.021	-0.013	0.055

Table 1: Parameter estimates for the RMN model on the margarine pricing dataset

To get some more insight in the effects of fluctuations in the price variable, we can calculate the corresponding average change in the probabilities of choosing each of the alternatives. In Table 2 we investigate the average effect on the probabilities of each of the alternatives if the price of one brand is increased by 10 cent. From this table we can see that, for example, increasing the price of Blue Bonnet leads to a decrease of 11.7% in its buying probability and at the same time increases the probability of buying each of the other brands. Note that these changes in probabilities sum to zero, as needed.

We calculated the same marginal changes in probability for every brand due to a 10 cent increase in price of one brand, but now using the non-robust multinomial probit model. The result of this analysis can be found in Table 3. For some of the brands, the predicted change in demand is not so different from the robust analysis. However, for other brands the difference is considerable. We consider two types of differences between both predictions.

	10 cent price increase for				
	Blue Bonnet	Fleisch-mann's	House Brand	Generic	Shed Spread
Δ Pr(Blue Bonnet)	-0.117	0.013	0.022	0.022	0.046
Δ Pr(Fleischmann's)	0.014	-0.047	0.009	0.008	0.017
Δ Pr(House Brand)	0.023	0.008	-0.078	0.014	0.029
Δ Pr(Generic)	0.023	0.008	0.014	-0.076	0.030
Δ Pr(Shed Spread)	0.057	0.018	0.033	0.032	-0.122

Table 2: Estimated marginal changes in probability by the robust RMN method when one brand's price is increased by 10 cent (for average values of the other variables)

The first type of difference is when the absolute value of the change in demand is different between the two approaches. For example, for Blue Bonnet, the robust method predicts a change in demand of 11.7%, while the non-robust method predicts a considerably smaller effect size, i.e. a 9.6% decrease. This indicates that a small amount of customers are very price insensitive and as a result the non-robust method underestimates the effect of the price change on the main group of customers.

	10 cent increase in price of				
	Blue Bonnet	Fleisch-mann's	House Brand	Generic	Shed Spread
Δ Pr(Blue Bonnet)	-0.096	0.012	0.018	0.019	0.038
Δ Pr(Fleischmann's)	0.013	-0.048	0.008	0.008	0.019
Δ Pr(House Brand)	0.018	0.008	-0.067	0.011	0.026
Δ Pr(Generic)	0.019	0.007	0.011	-0.066	0.025
Δ Pr(Shed Spread)	0.046	0.020	0.030	0.029	-0.108

Table 3: Estimated marginal changes in probability by the standard MNP method when one brand's price is increased by 10 cent (for average values of the other variables)

The second type of difference is when the predicted spread over the other brands is different for both methods. This is the case for Shed Spread, for example. The methods disagree on how the lapsed customers will spread over the remaining brands. While both methods indicate that most customers will switch to Blue Bonnet and to a lesser extent to

the House Brand and Generic brand, the robust results suggest that relatively more customers will switch to Blue Bonnet. Also, the robust method suggests that Fleischmann's would benefit relatively less from a price increase of Shed Spread than the non-robust approach indicates. Note that Fleischmann's is, on average, the most expensive brand, Blue Bonnet and Shead Spread are mid-range and the House Brand and Generic brand are, on average, the cheapest margarines. Both methods indicate that the cheapest and most expensive brands both have the lowest price sensitivity. However, the robust approach suggests a considerable higher price sensitivity for those brands. Again, this indicates that the behavior of some customers obscures the main effects in the data when the non-robust method is used.

We also investigated the predictive performance of the robust multinomial choice model and compared it with that of the standard multinomial probit model. To do so, we randomly split the margarine dataset in a training set and test set, containing 70% and 30% of the data respectively. The training data was used to estimate the model parameters for both models and these were then used to score the test set. The predictions were evaluated using three different measures. First, the percentage correctly classified (PCC) was calculated, with the predicted class set equal to the class with the highest predicted probability. PCC is a value between 0 and 1 with values closer to 1 signifying better predictive performance. Second, the prediction error (ERROR), $\sum_{i=1}^{n}\sum_{j=1}^{p}\sqrt{(cv_{ij}-cp_{ij})^2}$, was calculated. Here, cv is the $1 \times p$ vector containing a 0 when the option was not chosen and 1 if the option was chosen, while cp is the vector of choice probabilities. The closer the prediction error is to 0, the better the predictive performance of the model. Third, we calculated the area under the receiver operating curve for multiclass classification (mAUC) (Hand and Till, 2001). The resulting statistic is a value between 0.5 and 1 with higher values indicating better predictive performance. Finally, we repeated this procedure 5 times on different random subsets, resulting in a 5 times random cross validation.

	PCC	ERROR	mAUC
RMN	35.62	24.42	62.69
MNP	34.79	25.02	61.32

Table 4: Average percentage correctly classified, prediction error and multiclass AUC over 5 times random cross validation on the margarine dataset.

Table 4 shows the results of this analysis. All measures of predictive performance indicate that the proposed robust model has a slightly better predictive performance compared to the multinomial probit model. However, it is clear that the observed differences are rather limited on this dataset. It can be expected that the difference between RMN and MNP will be maximal when many outliers reside in the training data and none in the test data. Note that the prediction error was always better for the RMN model, while the other measures were less consistent. From this analysis, we can conclude that the predictive performance of the robust multinomial choice model is at least on par with that of the mulitnomial choice model, with a slight advantage for the RMN.

7 Discussion and conclusion

We introduced an outlier robust alternative to the popular multinomial logit and multinomial probit models. Similarly to the latter, our methodology allows for correlation between the regression equations and thus also alleviates concerns raised by IIA. Instead of using the logistic or multivariate normal distribution, our model uses the fat tailed multivariate Laplace distribution to model the latent utility differences. Moreover, a shrinkage procedure is applied on the multivariate observations to guarantee robustness toward outliers in both the dependent and independent variables.

Robust Bayesian estimation of the model parameters is conducted by MCMC. By exploiting the scale mixture of normals representation of the multivariate Laplace distribution, a Gibbs sampling algorithm can be set up instead of the more complex Metropolis-Hastings algorithm. This leads to a computationally efficient procedure to find the joint posterior distribution.

Monte Carlo experiments showed that our method captures well the effects present in the data, even when the data generating model deviates from the assumed model. The robust method focuses on the effects present in the majority of the data and is less influenced by deviating observations. It can handle outliers in the response as well as leverage points. We illustrated how our method can be used for pricing decisions using margarine scanner data.

Finally, it is important to realize that the proposed method also has a number of

limitations. First, while the multivariate Laplace distribution effectively reduces the effect of heavy tails in model (2), it is not yet known whether there is a distribution for the errors in model (1) that leads to a multivariate Laplace distribution for the error differences in (2). Such a result would strengthen further the proposed robust multinomial model. Second, the choice of the base alternative is not independent of the robustness properties of the model. That is, the outliers that appear in the base alternative are more likely to be shrunk towards the center of the data cloud. The reason is that in the differenced space these observations will become outliers in $p-1$ dimensions, while this is not the case for outliers in non-reference categories. Note that with outliers, the multinomial logit or probit model will also, and even more severely, be influenced by the choice of base alternative.

To provide robustness against leverage points, the weight function increases the uncertainty on model (2) for observations outside the high density region. For such observations, the relation becomes

$$u_i^\star = D_i\beta + (1/\omega_i)\varepsilon_i.$$

This can equivalently be written as

$$\tilde{u}_i^\star = D_i(\omega_i\beta) + \varepsilon_i,$$

which suggests that the robustness adjustment induces a non-linearity in the model. The effect of the covariates is reduced when they assume more extreme values. Consequently, our robust approach may mask non-linear effects in the choice model. It seems that modeling non-linearity or heterogeneity more generally in this type of robust Bayesian modeling will be very difficult, if not impossible. Note that in the context of statistical modeling the concept of what is an outlier depends highly on the model that is being considered. A sufficiently flexible model will always be able to accommodate the outliers as well, although this may not be appropriate because outliers are not assumed to be generated from a regular distribution.

The outlier shrinkage procedure uses a weight function based on robust Mahalanobis type distances as in (7). Hence, it is assumed that the high density region approximately has an elliptical shape. If the explanatory variables have more complex distributions such as a bimodal distribution for instance, then the high density region may not be

characterized well by a Mahalanobis type distance which implies a lesser performance of the RMN model. Moreover, finding a good value for α, the amount of shrinkage we want, is not straightforward. The optimal value depends on the shape of the multidimensional data cloud and might be difficult to assess. When the outliers form a relatively small cluster, the optimal value for α is the relative size of the cluster. However, when there is no clear boundary between the outliers and the bulk of the data, then relatively larger values for α might be more suitable. In case of bi-modal predictors, finding an optimal value for α becomes even more difficult.

Further research in this area is needed to alleviate the issues raised above. Moreover, the general ideas behind this Bayesian approach could also be extended to other models to diminish the effect of outlying observations. For example count data or panel data models could undoubtedly benefit from this methodology.

Acknowledgements

The computational resources (Stevin Supercomputer Infrastructure) and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by Ghent University, the Hercules Foundation and the Flemish Government. We thank editor Herman van Dijk and the anonymous reviewers for their constructive comments and helpful suggestions.

References

Agresti, A., 2002. *Categorical Data Analysis*. Wiley, New Jersey.

Allenby, G., Rossi, P., 1991. Quality Perceptions and Asymmetric Switching Between Brands. *Marketing Science*, **10**, 185-204.

Andrews, R.L., Currim, I.S., Leeflang, P.S.H., 2013. A Comparison of Sales Response Predictions From Demand Models Applied to Store-Level versus Panel Data. *Journal of Business & Economic Statistics,* **29**, 319–326.

Arslan, O., 2010. An alternative multivariate skew Laplace distribution: properties and estimation. *Statistical Papers*, **51**, 865–887.

Bayarri, C., Morales, J., 2003. Bayesian Measures of Surprise for Outlier Detection. *Journal of Statistical Planning and Inference*, **111**, 3–22.

Benoit, D.F., Van den Poel, D., 2012. Binary Quantile Regression: A Bayesian Approach Based on the Asymmetric Laplace density. *Journal of Applied Econometrics* **27**, 1174–1188.

Brown, B. M., 1983. Statistical Uses of the Spatial Median. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **45**, 25–30.

Burgette, L.F., Nordheim, E.V., 2012. The Trace Restriction: An Alternative Identification Strategy for the Bayesian Multinomial Probit Model. *Journal of Business & Economic Statistics* **30**, 404–410.

Chib, S., Greenberg, E., 1995. Understanding the Metropolis-Hastings algorithm. *American Statistician*, **49**, 327–335.

Čížek, P., 2012. Robust and Efficient Adaptive Estimation of Binary-Choice Regression Models. *Journal of the American Statistical Association*, **103**, 687–696.

Dunson, D.B., Watson, M., Taylor, J.A., 2003. Bayesian Latent Variable Models for Median Regression on Multiple Outcomes. *Biometrics*, **59**, 296–304.

Ebbes, P., Papies, D., van Heerde, H.J., 2013. The Sense and Non-Sense of Holdout Sample Validation in the Presence of Endogeneity. *Marketing Science,* **30**, 1115–1122.

Farias, V.F., Jagabathula, S., Shah, D., 2013. A Nonparametric Approach to Modeling Choice with Limited Data. *Management Science,* **59**, 305–322.

Geweke, J., 1991. Efficient Simulation from the Multivariate Normal and Student-t Distributions Subject to Linear Constraints. In *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface*, Keramidas, E.M. (ed). Interface Foundation of North America, Fairfax, VA, 571–578.

Geweke, J., 2001. Monte Carlo Simulation and Numerical Integration. In *Handbook of Computational Economics*, Amman, H.M., Kendrick, D.A., Rust, J. (eds). North-Holland: Amsterdam, 731–800.

Hand, D.J., Till, R.J., 2001. A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems. *Machine Learning,* **45**, 171–186.

Hausman, J.A., Abrevaya, J., Scott-Morton, F.M., 1998. Misclassification of the Dependent Variable in a Discrete-Response Setting. *Journal of Econometrics,* **87**, 239–269.

Huber, P., 1981. *Robust Statistics.* Wiley: New York, NY.

Kottas, A., Gelfand, A.E., 2001. Bayesian Semiparametric Median Regression Modeling *Journal of the American Statistical Association*, **96**, 1458–1468.

Kotz, S., Kozubowski, T., Podgorsky, K., 2001. *The Laplace Distribution and Generalizations: A Revisit with Applications to Communications, Economics, Engineering and Finance.* Birkhäuser: Boston, MA.

Lancaster, T., Jun, S.J., 2010. Bayesian Quantile Regression Methods, *Journal of Applied Econometrics*, **25**, 287–307.

Maddala, G.S., 1983. *Limited Dependent and Qualitative Variables in Econometrics.* Cambridge University Press: Cambridge, UK.

McCulloch, R., Rossi, P., 1994. An Exact Likelihood Analysis of the Multinomial Probit Model. *Journal of Econometrics*, **64**, 207–240.

Mebhane, W.R., Sekhon, J.S., 2004. Robust Estimation and Outlier Detection for Overdispersed Multinomial Models of Count Data. *American Journal of Political Science*, **48**, 392–411.

Molinari, F., 2008. Partial identification of probability distributions with misclassified data. *Journal of Eonometrics* **144**, 81–117.

Oja, H., 1983. Descriptive Statistics for Multivariate Distributions. *Statistics & Probability Letters* **1**, 327–332.

Peña, D., Zamar, R., Yan, G., 2009. Bayesian Likelihood Robustness in Linear Models. *Journal of Statistical Planning and Inference*, **139**, 2196–2207.

Ramalho, E.A., 2002. Regression models for choice-based samples with misclassification in the response variable. *Journal of Econometrics* **106**, 171–201.

Roelant, E., Van Aelst, S., 2007. An L1-type estimator of multivariate location and shape. *Statistical Methods & Applications*, **15**, 381–393.

Rossi, P., 2011. bayesm: Bayesian Inference for Marketing/Micro-econometrics. R package version 2.2-4. `http://CRAN.R-project.org/package=bayesm`

Rossi, P., Allenby, G., McCulloch, R., 2005. Bayesian Statistics and Marketing, John Wiley & Sons, New York.

Taddy, M.A., Kottas, A., 2010. A Bayesian Nonparametric Approach to Inference for Quantile Regression *Journal of Business & Economic Statistics*, **28**, 357–369.

Tanner, M.A., Wong, W.H., 1987. The Calculation of Posterior Distributions by Data Augmentation (with discussion). *Journal of the American Statistical Association*, **82**, 528–550.

Tierney, L., 1994. Markov Chains for Exploring Posterior Distributions. *Annals of Statistics*, **22**, 1701–1762.

Train, K., 1986. *Qualitative Choice Analysis: Theory Econometrics, and an Application to Automobile Demand.* MIT Press: Cambridge, MA.

Train, K., 2003. *Discrete Choice Methods with Simulation*, Cambridge University Press: Cambridge, UK.

Tukey, J., 1975. Mathematics and the picturing of data. In *Proceedings of the 1975 International Congress on Mathematics*, Vancouver, 523–531.

West, M., 1984. Outlier Models and Prior Distributions in Bayesian Linear Regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **46**, 431–439.

Yang, Y., He, X., 2012. Bayesian Empirical Likelihood for Quantile Regression, *Annals of Statistics*, **40**, 1102–1131.

Yu, K., Moyeed, R.A., 2001. Bayesian Quantile Regression. *Statistics & Probability Letters* **54**, 437–447.

Zellner, A., 1962. An Efficient Method of Estimating Seemingly Unrelated Regression Equations and Tests for Aggregation Bias. *Journal of the American Statistical Association,* **57**, 348–368.